# International Encyclopedia of STATISTICS

Edited by

WILLIAM H. KRUSKAL
*University of Chicago*

and

JUDITH M. TANUR
*State University of New York at Stony Brook*

VOLUME 2

THE FREE PRESS
*A Division of Macmillan Publishing Co., Inc.*
NEW YORK

Collier Macmillan Publishers
LONDON

esis has a privileged position. This is sometimes reasonable—for example, when the alternative hypotheses are diffuse while the null hypothesis is sharp. In other cases there is no particular reason to call one hypothesis "null" and the other "alternative" and hence no reason for asymmetry in the treatment of the two kinds of error. This symmetric treatment then is much the same as certain parts of the field called discriminant analysis. [*See* MULTIVARIATE ANALYSIS, *article on* CLASSIFICATION AND DISCRIMINATION.]

### Dangers, problems, and criticisms

Some dangers and problems of significance testing have already been touched on: failure to consider power, rigid misinterpretation of "accept" and "reject," serious invalidity of assumptions. Further dangers and problems are now discussed, along with related criticisms of significance testing.

[3]**Nonsignificance is often nonpublic.** Negative results are not so likely to reach publication as are positive ones. In most significance-testing situations a negative result is a result that is not statistically significant, and hence one sees in published papers and books many more statistically significant results than might be expected. Many—perhaps most—statistically nonsignificant results never see publication.

The effect of this is to change the interpretation of published significance tests in a way that is hard to analyze quantitatively. Suppose, to take a simple case, that some null hypothesis is investigated independently by a number of experimenters, all testing at the .05 level of significance. Suppose, further, that the null hypothesis is true. Then any one experimenter will have only a 5/100 chance of (misleadingly) finding statistical significance, but the chance that *at least one* experimenter will find statistical significance is appreciably higher. If, for example, there are six experimenters, a rejection of the null hypothesis by at least one of them will take place with probability .265, that is, more than one time out of four. If papers about experiments are much more likely to be published when a significance test shows a level of .05 (or less) than otherwise, then the nonpublication of nonsignificant results can lead to apparent contradictions and substantive controversy. If the null hypothesis is false, a similar analysis shows that the "power" of published significance tests may be appreciably higher than their nominal power. (Discussions of this problem are given in Sterling 1959; and Tullock 1959.)

[4]**Complete populations.** Another difficulty arises in the use of significance tests (or any other procedures of probabilistic inference) when the data consist of a complete census for the relevant population. For example, suppose that per capita income and per capita dollars spent on new automobiles are examined for the 50 states of the United States in 1964. The formal correlation coefficient may readily be computed and may have utility as a descriptive summary, but it would be highly questionable to use sampling theory of the kind discussed in this article to test the null hypothesis that the population correlation coefficient is zero, or is some other value. The difficulty is much more fundamental than that of nonnormality; it is hard to see how the 50 pairs of numbers can reasonably be regarded as a sample of any kind. Some statisticians believe that permutation tests may often be used meaningfully in such a context, but there is no consensus. [*For a definition of permutation tests, see* NONPARAMETRIC STATISTICS. *Further discussion of this problem and additional references are given in* Hirschi & Selvin 1967, chapter 13. *An early article is* Woofter 1933.]

**Target versus sampled populations.** Significance tests also share with all other kinds of inference from samples the difficulty that the population sampled from is usually more limited than the broader population for which an inference is desired. In the sleep-deprivation example the sampled population consists of students at a particular college who are willing to be experimental subjects. Presumably one wants to make inferences about a wider population: all students at the college, willing or not; all people of college age; perhaps all people. [*See* STATISTICS; ERRORS, *article on* NONSAMPLING ERRORS.]

**Neglect of power by word play.** A fallacious argument is that power and error of the second kind (accepting the null hypothesis when it is false) need not be of concern, since the null hypothesis is never really accepted but is just not rejected. This is arrant playing with words, since a significance test is fatuous unless there is a question with at least two possible answers in the background. Hence, both kinds of probabilities of wrong answers are important to consider. Recall that "accept" and "reject" are token words, each corresponding to a conclusion that is relatively more desirable when one or another true state of affairs obtains.

To see in another way why more than Type I error alone must be kept in mind, notice that one can, without any experiment or expense, achieve zero level of significance (no error of the first kind) by never rejecting the null hypothesis. Or one can achieve any desired significance level by using a random device (like dice or a roulette part of statistics called sequential analysis. Many sequential significance testing procedures have been carefully analyzed, although problems of determining reference sets nonetheless continue to exist. [*See* SEQUENTIAL ANALYSIS.]

**Optional stopping.** Closely related to the discussion of the preceding section is the problem of optional stopping. Suppose that an experimenter with extensive resources and a tendentious cast of mind undertakes a sequence of observations and from time to time carries out a significance test based on the observations at hand. For the usual models and tests, he will, sooner or later, reach statistical significance at any preassigned level, even if the null hypothesis is true. He might then stop taking observations and proclaim the statistical significance as if he had decided the sample size in advance. (The mathematical background of optional stopping is described in Robbins 1952.)

For the standard approach to significance testing, such optional stopping is as misleading and reprehensible as the suppression of unwanted observations. Even for an honest experimenter, if the sampling procedure is not firmly established in advance, a desire to have things turn out one way or another may unconsciously influence decisions about when to stop sampling. If the sampling procedure is firmly established in advance, then, at least in principle, characteristics of the significance test can be computed in advance; this is an important part of sequential analysis.

Optional stopping is, of course, relevant to modes of statistical analysis other than significance testing. It poses no problem for approaches to statistics that turn only on the observed likelihood, but many statisticians feel that these approaches are subject to other difficulties that are at least equally serious. [*See* BAYESIAN INFERENCE; LIKELIHOOD.]

**Simplicity and utility of hypotheses.** It is usually the case that a set of data will be more nearly in accord with a complicated hypothesis than with a simpler hypothesis that is a special case of the complicated one. For example, if the complicated hypothesis has several unspecified parameters whereas the simpler one specializes by taking some of the parameters at fixed values, a set of data will nearly always be better fit by the more complicated hypothesis than by the simpler one just because there are more parameters available for fitting: a point is usually farther away from a given line than from a given plane that includes the line; in the polynomial regression context, a linear regression function will almost never fit as well as a quadratic, a quadratic as well as a cubic, and so on.

Yet one often prefers a simpler hypothesis to a better-fitting more complicated one. This preference, which undoubtedly has deep psychological roots, poses a perennial problem for the philosophy of science. One way in which the problem is reflected in significance testing is in the traditional use of small significance levels. The null hypothesis is usually simpler than the alternatives, but one may be unwilling to abandon the null hypothesis unless the evidence against it is strong.

Hypotheses may be intrinsically comparable in ways other than simplicity. For example, one hypothesis may be more useful than another because it is more closely related to accepted hypotheses for related, but different, kinds of observations.

The theory of significance testing, however, takes no explicit account of the simplicity of hypotheses or of other aspects of their utility. A few steps have been made toward incorporating such considerations into statistical theory (see Anderson 1962), but the problem remains open.

**Importance of significance testing.** Significance testing is an important part of statistical theory and practice, but it is only one part, and there are other important ones. Because of the relative simplicity of its structure, significance testing has been overemphasized in some presentations of statistics, and as a result some students come mistakenly to feel that statistics is little else than significance testing.

**Other approaches to significance testing.** This article has been limited to the customary approach to significance testing based on the frequency concept of probability. For other concepts of probability, procedures analogous to significance testing have been considered. [*See* BAYESIAN INFERENCE. *An extensive discussion is given in* Edwards et al. 1963.] Anscombe (1963) has argued for a concept of significance testing in which only the null hypothesis, not the alternatives, plays a role.

WILLIAM H. KRUSKAL

[*See also* HYPOTHESIS TESTING.]

### BIBLIOGRAPHY

ANDERSON, T. W. 1962 The Choice of the Degree of a Polynomial Regression as a Multiple Decision Problem. *Annals of Mathematical Statistics* 33:255–265.

ANSCOMBE, F. J. 1963 Tests of Goodness of Fit. *Journal of the Royal Statistical Society* Series B 25:81–94.

BAKAN, DAVID 1966 The Test of Significance in Psychological Research. *Psychological Bulletin* 66:423–437.

BANCROFT, T. A. 1964 Analysis and Inference for Incompletely Specified Models Involving the Use of Preliminary Test(s) of Significance. *Biometrics* 20: 427–442.

BARNARD, G. A. 1947 The Meaning of a Significance Level. *Biometrika* 34:179–182.